# Exploiting Unlabeled Data for Feedback Efficient Human Preference-based Reinforcement Learning

Mudit Verma, Siddhant Bhambri, Subbarao Kambhampati
School of Computing & AI, Arizona State University
{muditverma, siddhantbhambri, rao} @ asu.edu

R2HCAI, AAAI 2023

## Unlabeled Trajectories & Representation Space

> **Observation 1:** An extremely large population of trajectories lie in the agent's buffer (collected over training episodes) that are not used in the reward learning process.

> **Observation 2:** the representation space for the reward function being learnt is not reflective of how the state space is structured.

> **Contribution:** We propose two corresponding loss functions that ensure participation of unlabeled trajectories in the reward learning process, and structure the embedding space of the reward model such that it reflects the structure of state space with respect to action distances.

## Human Preference-based Reinforcement Learning

- **Approximating the Human Reward Model**

> **Objective:** Compute probability for the human preferring trajectory $\tau_1$ over $\tau_2$:

$$P_\psi[\tau_0 \succ \tau_1] = \frac{\exp\left(\sum_t R_h(s_t^0, a_t^0)\right)}{\sum_{i\in\{0,1\}} \exp\left(\sum_t R_h(s_t^i, a_t^i)\right)}$$

> **Objective:** The human reward model can then by learnt by minimizing the cross-entropy between the predictions made by the supervised learner and the ground truth human labels as follows:

$$\mathcal{L}_{CE} = - \mathop{\mathbb{E}}_{(\tau_0,\tau_1,y)\sim\mathcal{D}} [y(0)\log P_\psi[\tau_0 \succ \tau_1] + y(1)\log P_\psi[\tau_1 \succ \tau_0]]$$

## Utilizing Unlabeled Data

- **Assumption 1: Preference on unlabeled trajectory**

> *A trajectory $\tau$, sampled under a policy $\pi_\phi$, that has not been queried to the human in the loop (HiL), is assumed to be preferred by the human.*

> Since there exist a large bank of trajectories that has not been queried to the HiL, Assumption 1 makes a paternalistic choice about whether those trajectories would be preferred by the HiL over some other trajectory.

> Moreover, we can use this assumption to ensure that the reward model can now use these unlabeled trajectories.

- **Proposed Triplet Loss**

> We propose a triplet loss that directly updates the reward model as follows:

$$\mathcal{L}^t(\tau; D_h) = \frac{1}{|D_h|} \sum_{\tau_g,\tau_b\sim D_h} max(0, ||\mathbf{R}(\tau) - \mathbf{R}(\tau_g)||^2 - ||\mathbf{R}(\tau) - \mathbf{R}(\tau_b)|| + m)$$

- **Definition 1: Preference on unlabeled trajectory**

> *Action distance $A_d$ between two states under some policy $\pi_\phi(s)$ and transition dynamics $T(s, a, s')$ is given by the expected number of action steps taken to reach a state $s'$ from $s$.*

> We propose to enforce such a soft constraint in the embedding space of the reward model, $R_e(s)$ computes the embedding of the state $s$, by ensuring that the Euclidean distance between the embedding of two states $s'$ and $s$ reflects the action distance $A_d(s, s')$.

- **Proposed Action Distance Loss**

> We propose an action distance loss that minimizes the Mean Squared Error (MSE) between the computed distance in the embedding space and the action distance as follows:

$$\mathcal{L}^a(D_p) = \frac{1}{|D_p|} \sum_{s_i,s_j,d_y\sim D_p} (||R_e(s_i) - R_e(s_j)||^2 - d_y)^2$$

Kambhampati's research is supported by the J.P. Morgan Faculty Research Award, ONR grants N00014-16-1-2892, N00014-18-1-2442, N00014-18-1-2840, N00014-9-1-2119, AFOSR grant FA9550-18-1-0067 and DARPA SAIL-ON grant W911NF19-2-0006.

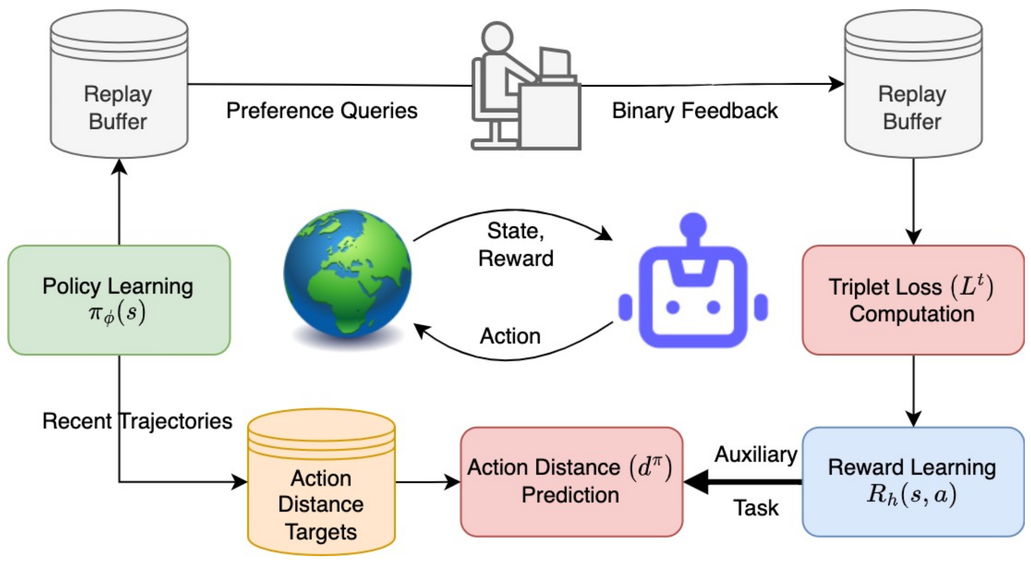



Figure 1: Overview of our proposed approach

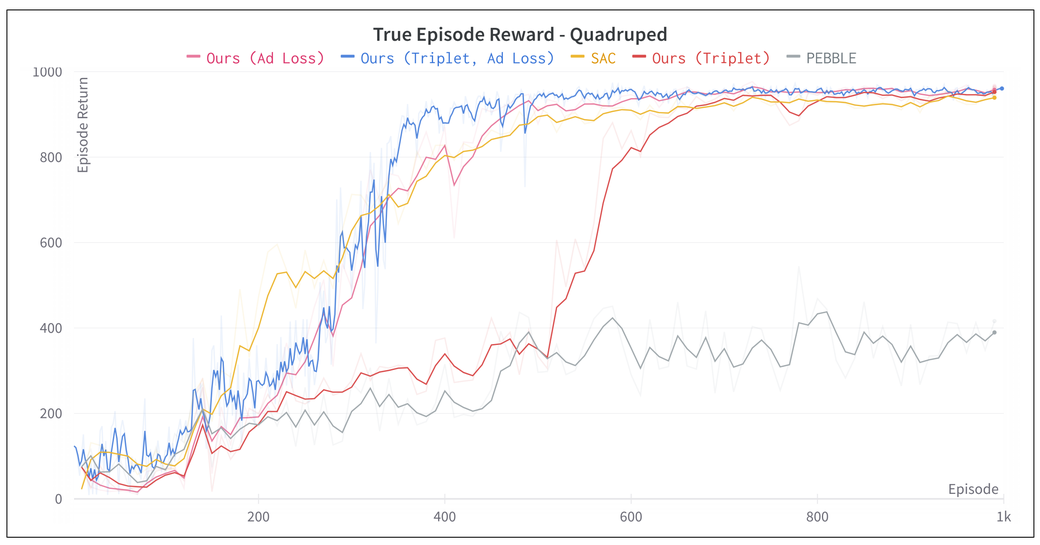
## Experiments & Results

- **Investigate via Empirical Evaluation**

> **Question 1:** Do the proposed losses improve the existing state of the art in preference-based RL in terms of reward recovery, feedback efficiency and performance of the learnt policy?

> **Question 2:** Are the two losses proposed in this work complementary, and more so synergic?


(a) Success Rate




(b) Return of learnt $\pi_\phi$ on ground truth reward $R_h$.

Figure 2: Evaluation curves on the robotic manipulation task of Sweep-Into and locomotion task of Quadruped-Walk

## Concluding Remarks

> We show that although these individual losses perform much better than the baseline PbRL and RL (SAC) in terms of reward recovery and human feedback sample efficiency, the synergic combi- nation of these yield a more powerful PbRL agent with low demands of human sample feedback and high performance.

>Future work includes a more thorough investigation of the effects of proposed method across diverse locomotion, robotic manipulation as well as explicit knowledge discrete domains.